Modeling Worldwide Gross Box Office Collection for Films 2012-2021

By: Rachel Goodridge

Abstract

Using data from films in 2012-2021 scraped from box office mojo, a model can be produced to determine which features affect the worldwide gross box office of a film. Features including budget, running time, and number of theaters show some positive correlation with worldwide gross. Comparisons can also be made between distributors and genres using dummy variables. Relative to Focus Features, Universal Pictures and Walt Disney Pictures perform better, while Open Road Films performs worse. Relative to Action, adding Animation, Music, and/or Musical aspects to a film would create improvement while Family, Fantasy, and Sports films would not.

Design

The motivation behind creating this model was to provide service to those participating in the film industry by being able to predict which factors will affect how much a movie will make in theaters worldwide. There can be many uses for this model including, but not limited to the following: Film makers may want to know which aspects of the process are crucially related to the overall yield. Actors may benefit from being able to make an informed decision about whether or not to participate in a movie. Investors can determine which films are worth putting their money into.

Data

The data is scraped from Box Office Mojo and contains information about various movies from 2012-2021. An individual sample is a singular movie (a row in the dataframe) and characteristics include Worldwide Gross, Distributor, Budget, MPAA, Running Time, Genres, Number of Theaters, Release Month, and Release Year (columns in the dataframe). The target for prediction in the model is Worldwide Gross.

Algorithms

First, I collected the target and features from a single movie on Box Office Mojo. Then, I expanded to scrape data from all movies within a single year and for all years between 2012-2021. After investigating dependent and independent variables, I performed a square root transformation on the target. Next, I built linear regression models based on various combinations of features with or without transformations and scored them using cross-validation. Finally, I compared the validation scores, chose the best model, and re-trained it on the entire training data set.

Tools

- requests, BeautifulSoup, and time for web scraping
- pandas, datetime, numpy, and pickle for data manipulation

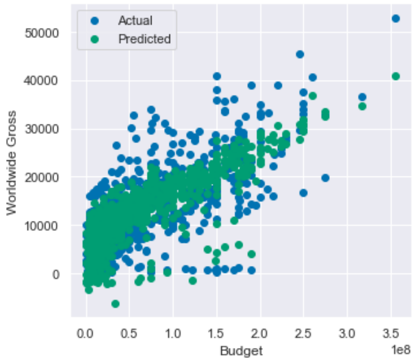
- statsmodels and sklearn for linear regression modeling
- matplotlib.pyplot, seaborn, and stats for plotting
- Python and Jupyter Notebook

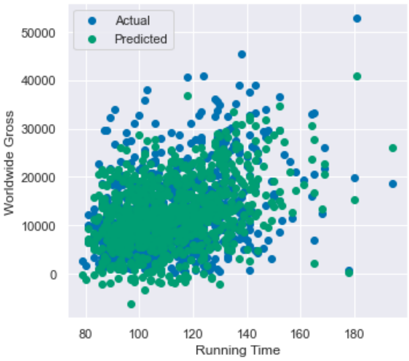
Communication

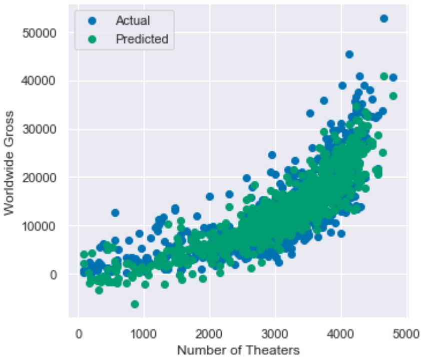
The final linear regression equation, including only features with significant p-values, is shown below. These numerical and dummy variables had meaningful enough correlation with the target variable (Worldwide Gross) to show some impact on the result. Increasing budget, running time, and number of theaters corresponds to increasing Worldwide Gross. For example, adding just one more theater would theoretically produce a $16.14 profit increase. Certain distributors like Universal Pictures and Walt Disney Pictures correlate with higher Worldwide Gross while Open Road Films corresponds with lower Worldwide Gross (compared to Focus Features as the baseline). For example, switching from Focus Features to Walt Disney Pictures would theoretically produce a $7,054,973 profit increase. Certain genres such as Animation, Music, and Musical may correspond to higher Worldwide Gross, while others such as Family, Fantasy, and Sport may be lower (compared to Action as the baseline). For example, animating a film would theoretically produce a $9,996,592 profit increase while creating a sports film would theoretically produce a $4,136,976 decrease in profit. Worldwide Gross has been lower these past couple years (likely due to COVID). 2020 and 2021 correlate with a decrease in Worldwide Gross compared to 2012 as the baseline. This can't really lead to predictions other than to say that it is better to produce films when we are not in a global pandemic.

$$\sqrt{WWGross} = -7857.75 + (.000056 \cdot Budget) + (40.9309 \cdot RunTime) + (4.01770 \cdot NumTheaters) - (1596.44 \cdot OpenRdFilms) + (1617.38 \cdot Universal) + (2656.12 \cdot Disney) + (3161.74 \cdot Animation) - (1701.24 \cdot Family) - (1233.66 \cdot Fantasy) + (2178.95 \cdot Music) + (1890.15 \cdot Musical) - (2033.96 \cdot Sport) - (4644.61 \cdot Year2020) - (3862.50 \cdot Year2021)$$

The fit of the model is decent (as seen in the figures below), but is likely missing some key features. Next steps would be to search for some more features such as production time/cost, actors/directions, and extent of film advertising to see how the model can be improved.


Actual
Predicted
Worldwide Gross
1e8
Budget


Actual
Predicted
Worldwide Gross
Running Time


Actual
Predicted
Worldwide Gross
Number of Theaters